Parsing de l'oral: traiter les disfluences

Marie-Laure Guénot

TALN 2005, Dourdan, 6–10 juin 2005

Parsing de l'oral : traiter les disfluences

Marie-Laure Guénot
Laboratoire Parole et Langage – CNRS / Université de Provence
{prenom.nom}@lpl.univ-aix.fr

Mots-clefs : Disfluences, Parsing, Linguistique de corpus, Linguistique formelle, Développement de grammaires, Grammaire de Construction (CxG), Grammaires de Propriétés (GP).

Keywords: *Disfluencies, Parsing, Corpus linguistics, Formal linguistics, Grammar development, Construction Grammar (CxG), Property Grammars (PG).*

Résumé Nous proposons une réflexion théorique sur la place d'un phénomène tel que celui des disfluences au sein d'une grammaire. Les descriptions fines qui en ont été données mènent à se demander quel statut accorder aux disfluences dans une théorie linguistique complète, tout en conservant une perspective globale de représentation, c'est-à-dire sans nuire à la cohérence et à l'homogénéité générale. Nous en introduisons une représentation formelle, à la suite de quoi nous proposons quelques mécanismes de parsing permettant de les traiter.

Abstract *We propose a theoretical reflexion about the place of a phenomenon like disfluencies, in a grammar. The precise descriptions that are available leads to a question : what status shall we give to disfluencies into a complete linguistic theory ?, keeping a global point of view and without compromising the coherence and the homogeneity of its representation. We introduce a formal representation of the phenomenon, and then we propose some parsing mechanisms in order to treat it.*

Introduction

On s'intéresse ici au traitement automatique des disfluences : phénomène non négligeable puisque très fréquent en oral spontané, la linguistique descriptive en a fourni un certain nombre d'études fines, présentant son organisation interne et ses caractéristiques. Cependant ces descriptions, quoique très précises dans leurs propositions, ne sont souvent pas exploitées en TALN, sans doute en partie parce que le statut des disfluences dans une grammaire n'y est pas défini de manière claire et formalisable. En effet les applications symboliques de traitement automatique qui s'efforcent d'analyser des données orales font appel à des techniques différentes pour traiter les disfluences, techniques qui sont pourtant basées pour la plupart sur les mêmes descriptions initiales.

Nous proposons ici une réflexion théorique concernant la place de phénomènes tels que les disfluences de l'oral dans une grammaire, laquelle grammaire a pour objet d'être représentée formellement, en vue notamment d'une exploitation en TALN. Nous conduisons notre réflexion en nous basant sur les travaux de linguistique descriptive, et dans le cadre du développement de plusieurs parseurs, aux caractéristiques et aux objectifs différents, mais qui sont tous basés sur une représentation formelle de descriptions linguistiques du français. Nous commencerons donc par exposer les études faites des disfluences en linguistique, puis nous montrerons comment nous avons interprété ces descriptions pour les rendre formalisables dans un modèle de représentation, avant de proposer un ensemble de mécanismes d'analyse automatique qui permettront d'exploiter cette grammaire et d'en tirer les résultats les plus efficaces possibles.

1 Situation du problème

1.1 Typologie(s) des disfluences

Dans la littérature linguistique, une disfluence est un endroit dans un énoncé où "*le déroulement syntagmatique est brisé*" (Blanche-Benveniste *et al.* (1990)) : on occupe une même place syntaxique avec plusieurs objets (ex. (1a)[1]).

(1) a. **il**
il a quand-même **un :**
une fibre pédagogique **assez :**
assez euh enfin réelle quoi

b. tu as toujours un rapport **(il) y a** **un directeur de fouilles**
(il) y a
(il) y a **les chouchous du directeur de fouille** et puis
les crétins de base enfin bon

Ce mécanisme n'est pas propre aux disfluences puisque l'on retrouve un même entassement dans les énumérations (ex. (1b)) ; en revanche, alors que dans ces dernières chaque occurrence de la même place ajoute un élément à la sémantique de l'énoncé, l'accumulation de *il* dans (1a) n'en modifie pas les caractéristiques sémantiques. Il serait de même abusif d'inscrire cette ré-

[1] Pour plus de clarté dans la lecture des disfluences, nos exemples sont représentés en grille (Blanche-Benveniste (1987)), et l'on notera **en caractères gras** les éléments illustratifs. Sauf mention contraire, tous les exemples de cet article sont tirés du Corpus d'Interactions Dilogiques (Bertrand & Priego-Valverde (2005)).

pétition comme étant une méthode de constitution syntagmatique (*i.e.*, l'accumulation paradigmatique ne forme pas de syntagme), ou de lui affecter des relations de dépendance syntaxique.

Parmi ces disfluences, on distingue deux grandes classes générales : les *bribes* qui sont des reprises à partir de syntagmes inachevés (ex. (1a)), et les *amorces* qui sont des reprises à partir de morphèmes inachevés (p. ex. *paran-* dans (2)).

(2) s'il n'y a pas d'éléments à mon avis euh il
il tombe **dans la paran-**
dans la parano quoi

Au sein des amorces Pallaud & Henry (2004) identifient trois formes différentes (formes que l'on peut, d'après elles, appliquer également aux bribes) : les amorces qui sont laissées *inachevées* (ex (3a)), celles qui sont *complétées* (ex. (2)), et celles qui sont *modifiées* (ex. (3b)).

(3) a. tu sais **j'ai v-** enfin
dans mon champ visuel (il) y a eu quelque chose tu vois

b. ils ont des ouvriers euh payés
spécialisés **sup-**
sur les chantiers de fouille

Pour sa part, Shriberg (1994), inspirée par Levelt (1983), a décrit l'organisation interne des disfluences en un ensemble d'espaces distincts : le *reparandum* qui est le lieu de la première production, inachevée au niveau du point d'interruption (*interruption point*), suivi de l'*interregnum* au sein duquel il peut se produire soit rien, soit une marque d'hésitation, soit une à plusieurs nouvelles tentatives de formulation (inachevées), jusqu'au *repair* qui correspond à la reprise du déroulement syntagmatique.

Toutes ces études, qui décrivent l'organisation interne des disfluences, peuvent être prises en considération dans le développement d'une formalisation. Elles les présentent comme un phénomène unique, avec des caractéristiques régulières (l'entassement paradigmatique, l'absence de fonction syntaxique et de fonction sémantique, les espaces internes), et des caractéristiques plus spécifiques à certains cas (bribes *vs.* amorces, inachèvement *vs.* complétion *vs.* modification, composition de l'interregnum). Cependant elles n'indiquent pas comment l'on différencie une disfluence d'une autre construction, ni comment l'on doit les traiter lors de l'analyse d'un énoncé.

1.2 Les disfluences en TALN

Observons maintenant comment le phénomène est traité en TALN[2] : en dépit de la prise en considération, dans la plupart des cas, de tout ou partie des descriptions exposées ci-dessus, les solutions concrètes proposées pour le traitement automatique sont nettement différentes suivant la tâche à accomplir et le type d'approche.

La première technique que l'on peut rencontrer consiste à "effacer" les disfluences de l'entrée qui sera analysée, en effectuant un pré-traitement des données dont l'objet est de reconnaître les disfluences et de les remplacer par une forme considérée comme "équivalente" ne présentant pas de rupture du déroulement syntagmatique (p.ex. chez Dowding *et al.* (1999)). On peut se

[2]Parce que l'on se place nous-même dans la perspective générale de la représentation formelle de la langue, on ne s'intéresse ici qu'aux méthodes de TALN qui sont basées sur des descriptions linguistiques, et non sur les techniques probabilistes, qui certes proposent des approches intéressantes, mais ne font pas partie de notre cadre de recherche.

demander quel est précisément le niveau d'"équivalence" recherché, et quelles sont les limites imposées par des résultats d'analyse pour des utilisations ultérieures, si ceux-ci sont basés sur des entrées qui ne contiennent plus la totalité des informations linguistiques produites.

La deuxième technique consiste en quelque sorte à "ignorer" les disfluences, *i.e.* à ne pas les prendre en compte lors de l'analyse (cf. par exemple Pérennou (1996)). Ceci permet d'obtenir un résultat de parsing dit "robuste", mais ne pose pas la question du statut des unités qui n'ont pas été considérées dans l'analyse : bien que les disfluences n'aient pas de fonction syntaxique en tant que telles, chaque élément qui occupe une place syntagmatique remplit en lui-même la fonction syntaxique de cette place, et il semble difficile d'admettre dans ce cas que seule une occurrence de chaque place sera considérée dans l'analyse. En d'autres termes, si l'on n'analyse que le *repair*, quel est le statut des constituants des autres espaces de la disfluence ?

La troisième technique est celle qui consiste à regrouper les disfluences en un groupe. Antoine *et al.* (2003) proposent dans cette perspective des analyseurs qui forment des disfluences en rassemblant des chunks (chacun d'eux devant être une occurrence de la même place syntaxique) en vertu de "*relations de dépendance sémantico-pragmatiques*". Dans un même ordre d'idées, Godfrey *et al.* (1992) proposent une méthode d'annotation des disfluences qui consiste à effectuer un parenthésage de la totalité de l'accumulation paradigmatique. Dans ce cas on peut se demander comment est déterminée la catégorie syntagmatique de cet ensemble (indispensable à l'analyse), qui peut être constitué de plusieurs répétitions ne comptant pas toujours les mêmes constituants, et qui ne correspond pas systématiquement à un syntagme complet.

On voit que les techniques qui prennent en compte les disfluences, bien que se basant sur tout ou partie des descriptions données ci-dessus, diffèrent nettement dans leurs traitements, qui sont tous limités par la nature de leur représentation du phénomène. Nous allons donc commencer par expliquer quelle place nous lui donnons dans notre grammaire, avant d'en montrer la formalisation proposée à partir de là.

2 Description et représentation

Le fait de vouloir intégrer un phénomène tel que celui des disfluences dans une grammaire, met en avant une différence fondamentale entre les descriptions linguistiques et leur formalisation : là où la première s'attache à décrire finement le fonctionnement interne et les propriétés d'un phénomène donné, le modèle formel qui a pour tâche de le représenter doit également intégrer, en plus de sa description interne, les propriétés plus générales du phénomène, *i.e.* conserver un point de vue plus global sur l'articulation entre celui-ci et les autres, sur la façon dont le tout s'articule et conserve une cohérence générale.

C'est ce à quoi nous allons nous attacher ici. Nous allons proposer une réflexion générale sur la place des disfluences au sein d'un système grammatical. La grammaire du français que nous développons et dans laquelle nous tâchons d'intégrer ce que nous présentons ici a pour cadre théorique celui de la *Construction Grammar* (CxG, cf. p.ex. Kay & Fillmore (1999)), et pour modèle formel celui des *Grammaires de Propriétés* (GP, cf. p.ex. Blache (2005)). La description et la représentation formelle qui suivent sont donc basées sur ce cadre de travail CxG/GP, cependant nous espérons les présenter de telle sorte qu'elles puissent s'appliquer à d'autres cadres théoriques et formels que celui qui fait l'objet de notre travail.

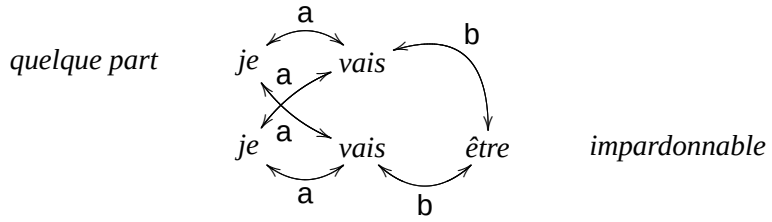
Ce que l'on a pu voir jusqu'ici met en avant le fait que pour traiter des disfluences, il est né-

cessaire de commencer par répondre à une question d'ordre plus général : *Quelle est leur place dans la grammaire ?* Et pour répondre à cette question, la première chose que l'on doit se demander est ce que l'on représente au juste dans une grammaire : quels sont les objets que l'on y manipule ? Y représente-t-on des relations entre les occurrences possibles d'un énoncé, ou alors entre les places syntaxiques occupées par ces occurrences ? Dans le cas d'énoncés sans disfluences, ces questions ne sont pas si évidentes ; ainsi, dans un énoncé tel que celui de la figure 1[3], les *occurrences possibles* et les *places syntaxiques* sont confondues, puisque chaque place syntaxique n'est occupée que par une unique occurrence.

(1)

FIG. 1 – Relations dans un énoncé sans disfluences.

Par contre, quand on traite des énoncés avec des disfluences, ces questions prennent toute leur importance. Considérons d'abord que l'on représente dans une grammaire des *relations entre des occurrences possibles.* L'analyse d'une disfluence peut alors être représentée selon l'illustration de la figure 2.



(2)

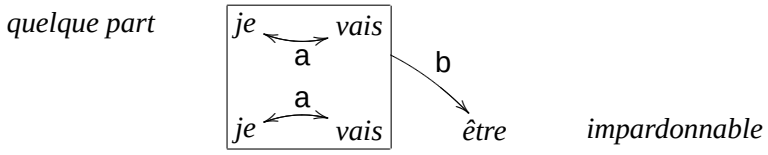
FIG. 2 – Relations entre occurrences.

On voit que dans ce cas on multiplie le nombre de chaque relation par le nombre d'occurrences de la même place syntagmatique : la relation a figure quatre fois et la b deux fois, au lieu d'une seule dans la figure 1. La conséquence de cela est que l'ensemble de caractéristiques, spécifique au syntagme qui contient la disfluence (ici, *je vais je vais être impardonnable*), varie non seulement en fonction de la présence, mais aussi de la forme d'une disfluence : l'ensemble {a, b} de la figure 1 devient {a, a, a, a, b, b} pour la disfluence précise de la figure 2. De plus, un certain nombre de propriétés définitoires du syntagme (*e.g.* l'unicité du pronom clitique nominatif, ou l'ordre linéaire entre ce même pronom et le verbe) sont faussées par la présence de la répétition, et la définition dans la grammaire doit tenir compte de ces variations de caractéristiques. Pourtant comme on l'a vu plus haut, les caractéristiques spécifiques à la présence d'une disfluence n'ont pas d'incidence sur l'analyse syntaxique d'un énoncé, et donc ne devraient pas avoir d'incidence sur la définition (syntaxique) d'un syntagme.

Représenter des relations entre occurrences ne semble donc pas être le fait d'une grammaire ; considérons alors que l'on y représente des *relations entre des places syntagmatiques.* Dans ce cas, si l'on suit l'approche de Godfrey *et al.* (1992), on peut illustrer le traitement des disfluences comme dans la figure 3.

Ici l'on résout le problème de la multiplication injustifiée des caractéristiques, mais on se trouve face à un autre problème : pour pouvoir faire une analyse syntaxique il faut qu'à chaque place

[3]Les "relations" que l'on représente ici sont des représentations intuitives, et n'illustrent pas une théorie donnée.



(3)

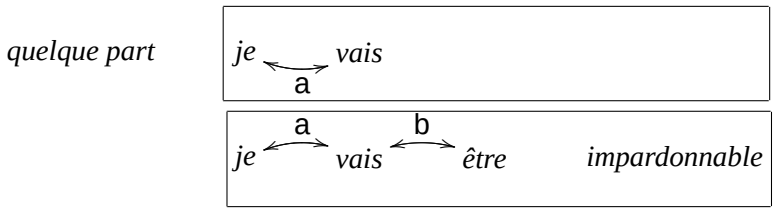
FIG. 3 – Relations entre places syntagmatiques.

corresponde une catégorie, et dans l'exemple quelle est la catégorie de *je vais je vais*, ou de *c'était je crois qu'il était* dans l'ex. (4) ?

(4) **c'était**
je crois qu'il était autrichien ou un truc comme ça

En effet, s'il est simple de traiter de cette façon les reprises simples telle que *il il* dans l'exemple (1a) en lui affectant la place de "pronom clitique", il devient plus difficile de s'accorder sur le statut d'un groupe constitué d'un fragment de début de syntagme, qui ne correspond à aucune étiquette syntaxique. Comment intégrer dans une analyse syntaxique, des objets qui ne sont pas des éléments syntaxiques ? Il faudrait pour cela les ajouter artificiellement à la grammaire, intégrer ces groupes qui ne sont pas vraiment des syntagmes, mais dont la seule raison d'y figurer est qu'ils peuvent apparaître en tant qu'occurrence. Au-delà du problème linguistique de fond que ce type d'artefact suppose (quelle analyse fait-on ? quelle est la nature des objets que l'on introduit ?), on en revient également au problème posé par la technique précédente : les disfluences ne sont pas décrites en tant que phénomène, mais chacune des possibilités de disfluence devra être l'objet d'une construction particulière, et l'on sera limité à un moment ou à un autre par l'itération limitative des possibilités *a priori* infinies.

Une façon de remédier à ce problème sans fabriquer de catégories *ad hoc* est de ne pas rassembler les différentes occurrences d'une disfluence en un groupe unique, mais de considérer que chaque reprise est une occurrence, achevée ou non, du syntagme complet (figure 4).



(4)

FIG. 4 – Les disfluences comme occurrences complètes de syntagmes.

Pour ne pas recourir à des catégories vides[4], on peut tout simplement considérer (ce qui est tout à fait cohérent dans le cadre CxG/GP) que la caractérisation des occurences inachevées de syntagme seront le reflet de leur constitution : un certain nombre de propriétés seront, à juste titre, non évaluées, et d'autres seront évaluées et non-satisfaites, en comparaison avec le *repair*. Il s'agit ensuite, pour ne pas se contenter de déplacer au niveau supérieur le problème posé par la première possibilité envisagée ici, de mettre en relation ces occurrences du même syntagme au sein de la grammaire, en tant que "phénomène de disfluence". Il est possible de le repérer par un ensemble de caractéristiques telles que la différence de caractérisation entre les premières occurrences, incomplètes, et le *repair*, ainsi que l'occupation des mêmes fonctions syntaxique et sémantique. D'autres éléments peuvent ensuite permettre de distinguer les différentes formes

[4]Nous n'avons pas la place de justifier notre position sur ce point ici, cependant nous développons une grammaire qui n'a pas de recours au postulat de catégories vides, que ce soit pour ce cas ou pour tous les autres.

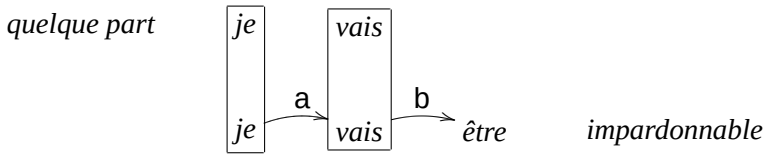
de disfluences : les *inachevées* n'auront pas les mêmes constituants que le repair, les *modifiées* s'en écarteront par un sous-ensemble variable de traits (différences de genre, de nombre, etc.), contrairement aux *complétées* dont les constituants (présents) seront identiques à ceux du repair.

Ce qui pose un problème plus délicat dans cette représentation, c'est l'expression de relations entre parties différentes d'occurrences différentes. Observons par exemple l'énoncé suivant :

(5)
lesquels registres	sont très	euh
	sont	
doivent	être	
doivent pouvoir	être	contrôlés

tiré de Blanche-Benveniste *et al.* (1990), p. 24, où l'auteur dit que l'on "*vise à dégager la séquence maximale qui a été donnée par le locuteur, en tenant compte de toutes les bribes qu'il a fournies ; dans l'exemple précédent, on retiendra comme séquence maximale :* lesquels registres doivent pouvoir être très contrôlés", ce qui implique de pouvoir établir des relations entre *très* apparu dans le reparandum, et *contrôlés* apparu dans le repair. Or, bien qu'une analyse en GP repérera cette relation, il paraît par contre bien difficile de la représenter au sein même de la grammaire et donc d'en tenir compte dans la description de la "*séquence maximale*". Difficile, parce que cela demanderait d'utiliser une relation entre constituants de syntagmes différents, lorsque ces syntagmes sont liés entre eux par une relation de disfluence, et que l'on n'utilise pas de telles relations dans d'autres cas. On devrait donc créer artificiellement, non pas de nouveaux syntagmes comme précédemment, mais de nouvelles propriétés (ou de nouvelles *sémantiques* de propriétés, cf. Vanrullen *et al.* (2003)) uniquement pour traiter le phénomène.

La représentation que l'on propose conserve la vision des disfluences en tant que construction à part entière, telle que montrée ci-dessus, mais ne la considère pas systématiquement comme mettant en relation des syntagmes : on reliera en tant que disfluence chaque occurrence répétée d'une même place syntaxique, qu'elle soit construite ou non (figure 5).



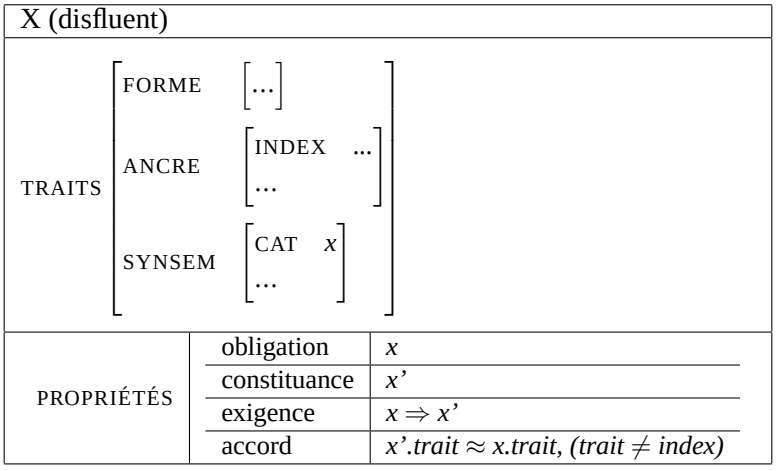
(5)

FIG. 5 – Les disfluences comme accumulation de places syntagmatiques.

Ainsi, toutes les places syntaxiques qui doivent être mises en relation lors de l'introduction du syntagme qui les contient (le "syntagme maximal" de Blanche) sont séparées les unes des autres et peuvent être prises en compte par le reste de la grammaire de la même façon, qu'elles soient disfluentes ou non, sans introduire ni formes de constructions ni relations *ad hoc*. Concrètement, une bribe *complétée* aura la forme de la figure 4, et une *modifiée* aura cette même forme à ceci près que les éléments d'une disfluence auront quelques différences de valeurs de traits. Les *inachevées*, enfin, seront traitées comme des reprises de syntagmes (donc comme décrites dans le cas précédent, mais uniquement quand aucune identité de forme ne permet de relier des unités plus petites entre elles).

Dans la grammaire, la construction correspondant aux disfluences aura la forme proposée dans la figure 6[5]. Ces quelques caractéristiques nous permettent de signifier qu'une disfluence consiste en la mise en relation d'un *x* et un ou plusieurs *x'*, dont chacun des traits, mis à part l'INDEX,

[5]Il s'agit d'une représentation extrêmement simplifiée pour les besoins de l'article. La développer ici sous sa forme réelle eût demandé des explications superflues dans ce contexte précis.



(6)

FIG. 6 – représentation formelle de la construction “disfluente”.

sont de valeurs identiques un à un. La satisfaction complète de ces propriétés caractérise une bribe *complétée* ; si un certain nombre de propriétés d’accord sont évaluées et non satisfaites alors on caractérise une bribe *modifiée*, et l’application de *x* et *x’* en tant que syntagmes nous permet de reconnaître les bribes *inachevées*. Les *accords*, dans notre grammaire, font référence à la totalité des traits d’un objet, que ceux-ci soient morphologiques ou syntaxiques ou sémantiques (ou autres), ce qui nous permet de traiter avec une même description tout aussi bien les disfluences dont la modification est syntaxique (*un une* dans (1a)), que celles dont la modification est sémantique (ex. (6)).

(6) ils sont pas à l’abri de ça quoi mais c’est **un peu**
pas mal d’hypocrisie quand-même à ce niveau-là

3 Mécanismes de parsing

Dans notre cadre on s’intéresse principalement à des tâches de parsing non-déterministe, cependant l’on suppose (et l’on espère) que les quelques mécanismes qui suivent pourront s’appliquer aussi bien à du parsing déterministe. En outre, l’implémentation de cette grammaire est en cours, mais à l’heure actuelle elle n’est pas suffisamment avancée pour que l’on puisse la considérer comme évaluable. Nous verrons donc ici les idées générales qui dirigent la phase d’implémentation qui est en cours de réalisation.

Nous avons vu jusque là comment représenter les disfluences au sein de la grammaire, et pourquoi ; voyons maintenant quelles conséquences cette introduction peut avoir sur le parsing. En effet, les tenants et les aboutissants de l’automatisation d’une analyse basée sur une grammaire formelle ont ceci de différent du développement de la grammaire elle-même, qu’ils doivent fournir un résultat exploitable à l’issue d’un traitement le plus efficace et le plus robuste possible. L’obstacle principal ici est un problème d’explosion combinatoire : la description des disfluences proposée est tellement large qu’elle va engendrer l’introduction de constructions disfluentes non seulement dans les cas pertinents, mais également dans une quantité déraisonnable de cas superflus. Plusieurs façons de limiter les introductions superflues sont envisageables :

– Borner l’introduction d’une disfluence à une distance arbitrairement définie (qu’elle soit fixe ou relative à la longueur de l’énoncé). Même si les disfluences peuvent être non bornées (et elles le sont dans de nombreux cas selon notre description), elles ne sont probablement que

très rarement séparées de plus d'une certaine distance.
– Introduire une série de marques linguistiques permettant de différencier les disfluences et les énumérations (c'est ce que font p.ex. Johnson *et al.* (2004)). Les différentes parties des disfluences auraient tendance à être séparées par des pauses oralisées, des connecteurs, alors que les énumérations seraient plutôt séparées par des coordonnants. De plus, les énumérations sont des entassements de syntagmes dont des occurrences sont (normalement) toutes achevées, contrairement aux disfluences comme on l'a vu.

A terme, l'observation des résultats fournis en parsing en faisant varier ces différentes possibilités devrait nous permettre de faire remonter des informations exactes à ce propos, que l'on pourra intégrer directement à la grammaire comme autant de propriétés supplémentaires des disfluences.

Un autre mécanisme à traiter lors du parsing est celui de l'instanciation des traits de la construction disfluente en vertu des traits de ses constituants. Notre méthode consisterait, pour le cas où l'accord n'est pas satisfait, à affecter à la construction disfluente la valeur de trait de l'occurrence du repair, pour justifier de l'analyse par exemple de *un une* dans (1a) comme étant un déterminant de genre féminin (et non simplement indéterminé). Ceci permettrait d'affiner l'analyse et donc de réduire la possibilité d'introduction de relations non pertinentes par la suite.

Conclusion et perspectives

Les disfluences dont nous traitons ici ont été décrites comme des entassements paradigmatiques qui ont ceci de particulier qu'ils n'ont ni de fonction syntaxique, ni de fonction sémantique. Les objets accumulés sur une même place syntaxique peuvent être parfaitement identiques ou partiellement différents, mais ont toujours un certain nombre de caractéristiques communes, et l'on peut en décrire une organisation interne assez précise. Cependant l'exploitation très variable que l'on peut en voir en TALN montre qu'en plus de tout cela, il est nécessaire avant de traiter les disfluences dans un modèle basé sur un formalisme linguistique, de répondre aux questions suivantes : Quelle est la place des disfluences dans une grammaire ? Quand et comment les analyse-t-on ? Nous avons donc ici proposé une réflexion sur la place des disfluences dans une grammaire formelle, à la suite de laquelle nous avons introduit notre représentation de ce phénomène. Au delà du problème posé par cette représentation particulière, nous avons mené une réflexion plus générale sur la place de ce type de phénomènes, propres à l'oral, et par là nous avons présenté une vision globale du développement de grammaire et de ces spécificités. Nous avons proposé ensuite quelques mécanismes de traitement de la grammaire proposée, qui devraient permettre de garantir des résultats plus pertinents et robustes pour la tâche précise du parsing, tout en mettant en avant la méthode de développement de grammaire assisté par l'informatique, qui permet d'effectuer un va-et-vient entre les hypothèses formées à partir de l'étude de corpus, leur formalisation et leur vérification sur une quantité importante de données de manière automatisée.

Ce travail ne s'arrête bien évidemment pas là. Nous avons présenté ici une étude dont l'implémentation est en cours, qui a pour objectif à terme de proposer une grammaire formelle du français oral, basée sur des descriptions fines de corpus. Un autre objectif qui transparaît à travers cet article, et qui sera sans aucun doute nécessaire à la finalisation de cette grammaire, est l'intégration au sein même de la grammaire formelle, d'informations de plusieurs domaines différents : nous avons évoqué la syntaxe et la sémantique ici, mais pour traiter de l'oral il nous semble évident qu'à cela nous devrons ajouter des informations prosodiques (p. ex. Morel &

Danon-Boileau (1998)). Nous avons également pour projet d'ajouter à cela les informations gestuelles pertinentes[6], suite à l'étude du corpus qui nous a servi de base pour cet article, et qui est disponible sous forme audiovisuelle. Il serait également intéressant d'intégrer au traitement des éléments concernant l'interprétation (pragmatique et/ou psycholinguistique) de ces disfluences, et qui justifient leur apparition et leur statut lors de la perception d'un message. Enfin, outre le développement de grammaire, cette étude s'inscrit dans un projet d'annotation multi-niveaux de corpus, et dans ce cadre permet de réfléchir aux différences et aux liens existants entre les différents niveaux à représenter, en se basant sur l'étude de phénomènes réels et concrets.

Références

Jean-Yves Antoine, Jérome Goulian, & Jeanne Villaneau. Quand le tal robuste s'attaque au langage parlé : analyse incrémentale pour la compréhension de la parole spontanée. In *Actes de TALN 2003*, 2003.

Roxanne Bertrand & Béatrice Priego-Valverde. Le corpus d'interactions dilogiques : Présentation et perspectives. Technical report, Laboratoire Parole et Langage – CNRS / Université de Provence, 2005.

Philippe Blache. Property grammars : A fully constraint-based theory. In H Christiansen, P Skadhauge, & J Villadsen, editors, *Constraint Satisfaction and Language Processing*. Springer-Verlag, 2005.

Claire Blanche-Benveniste. Syntaxe, choix du lexique et lieux de bafouillage. *DRLAV*, 36-37 :123–157, 1987.

Claire Blanche-Benveniste, Mireille Bilger, Christine Rouget, & Karel Van Den Eynde. *Le français parlé : Etudes grammaticales*. Sciences du langage. CNRS Editions, Paris, 1990.

J. Dowding, J. Gawron, D. Appelt, J. Bear, L. Cherny, R. Moore, & D. Moran. Gemini : A natural language system for spoken language understanding. In *Proceedings of ARPA Workshop on Human Language Technology*, pages 21–24, 1999.

J. J. Godfrey, E. C. Holliman, & J. McDaniel. Switchboard : A telephone speech corpus for research and development. In *Proceedings of the IEEE*, pages 517–520, 1992.

Marie-Laure Guénot & Emmanuel Bellengier. Quelques principes pour une grammaire multimodale du français. In *Proceedings of RECITAL 2004*, 2004.

Mark Johnson, Eugene Charniak, & Matthew Lease. An improved model for recognizing disfluencies in conversational speech. In *Rich Transcription 2004 Fall workshop*, 2004.

Paul Kay & Charles J. Fillmore. Grammatical constructions and linguistic generalizations : The *what's X doing Y ?* construction. *Language*, 75(1) :1–33, March 1999.

W Levelt. Monitoring and self-repair in speech. *Cognition*, 14 :41–104, 1983.

Mary-Annick Morel & Laurent Danon-Boileau. *Grammaire de l'intonation*. Bibliothèque de faits de langues. Ophrys, Paris, 1998.

Berthille Pallaud & Sandrine Henry. Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé. In *Actes de JADT 2004*, 2004.

G. Pérennou. Compréhension du dialogue oral : le role du lexique dans l'approche par segments conceptuels. In *Actes de Lexique et Communication Parlée*, pages 169–178, 1996.

Elizabeth Shriberg. *Preliminaries to a theory of speech dislfuencies*. PhD thesis, University of California, Berkeley, 1994.

Tristan Vanrullen, Marie-Laure Guénot, & Emmanuel Bellengier. Formal representation of property grammars. In *Proceedings of ESSLLI Student Session*, 2003.

[6]Ce point avait été introduit dans Guénot & Bellengier (2004).